GC – Integrated Web Environment for Corpus Linguistics

What is GC?

GC is a Web tool being developed at Linguateca/CLUP that aims to provide a comprehensive work environment for Corpus-Based Linguistic Research. GC allows users to:

- access several Corpora tools from a single entry point using a regular web browser
- access and query generic Corpora (BNC, Reuter's, COMPARA, CETEMPúblico)
- build personal simple, parallel and comparable Corpora from text files (PDF, PS, Word, HTML, TXT)
- use several (on-line/off-line) tools with their personal Corpora (statistics, POS-taggers, Filters, etc.)
- communicate and exchange results with other users

Motivation

- Lack of Comprehensive, wide-scope Corpora Tools
- Commercial Packages are usually difficult to Integrate/Customize
- Tools are not prepared to support cooperative work.
- Linguistic knowledge is not usually integrated in tools.

Internet Integration

GC provides seamless integration with the World Wide Web allowing users to:

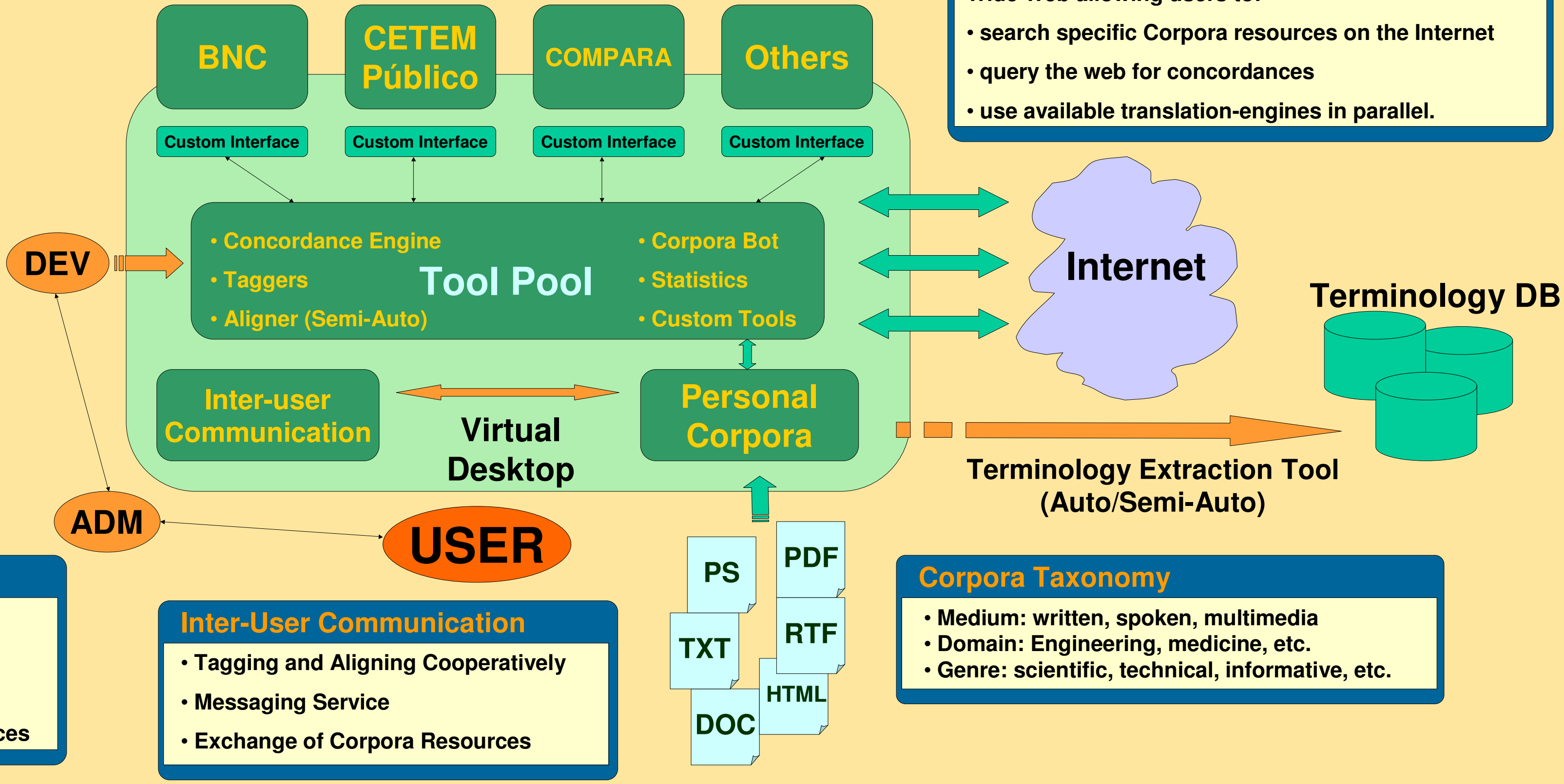
- search specific Corpora resources on the Internet
- query the web for concordances
- use available translation-engines in parallel.

Developer's Tasks:

- Integrate Existing Tools/Resources
- Develop Additional Generic Tools
- Interact with Users/Administrator
- Develop Custom Tools for particular research needs

Administrator's Tasks:

- Users, Groups and Disk Quotas
- Corpora Taxonomy (see box)
- Documentation Organization
- Access Service Statistics



BNC

CETEM Público

COMPARA

Others

Custom Interface

Custom Interface

Custom Interface

Custom Interface

- Concordance Engine
- Taggers
- Aligner (Semi-Auto)

Tool Pool

- Corpora Bot
- Statistics
- Custom Tools

Inter-user Communication

Virtual Desktop

Personal Corpora

DEV

ADM

USER

Internet

Terminology DB

Terminology Extraction Tool (Auto/Semi-Auto)

PS

PDF

TXT

RTF

HTML

DOC

Teacher's Tasks:

- Provide on-line tutorials
- Provide links to:
  - on-line teaching material
  - bibliography and other resources

Inter-User Communication

- Tagging and Aligning Cooperatively
- Messaging Service
- Exchange of Corpora Resources

Corpora Taxonomy

- Medium: written, spoken, multimedia
- Domain: Engineering, medicine, etc.
- Genre: scientific, technical, informative, etc.

FLUP/CLUP
http://www.letras.up.pt

LINGUATECA
http://www.linguateca.pt

Belinda Maia [FLUP/CLUP] & Luís Sarmento [Linguateca@CLUP]
bmaia@mail.telepac.pt las@letras.up.pt